## Title

Determining cancer risk: the evolutionary multistage model or total stem cell divisions?

**Permalink**
https://escholarship.org/uc/item/0b60p06c

**Journal**
Proceedings of the Royal Society B, 287(1941)

**ISSN**
0962-8452

**Authors**
Nunney, Leonard
Thai, Kevin

**Publication Date**
2020-12-23

**DOI**
10.1098/rspb.2020.2291

Peer reviewed

PROCEEDINGS B

royalsocietypublishing.org/journal/rspb

Research

**Cite this article:** Nunney L, Thai K. 2020 Determining cancer risk: the evolutionary multistage model or total stem cell divisions? *Proc. R. Soc. B* **287**: 20202291. https://doi.org/10.1098/rspb.2020.2291

Received: 15 September 2020
Accepted: 24 November 2020

**Subject Category:**
Evolution

**Subject Areas:**
health and disease and epidemiology, evolution

**Keywords:**
cancer, multistage carcinogenesis, multistep carcinogenesis, lifetime stem cell divisions, Peto's paradox, evolution

**Author for correspondence:**
Leonard Nunney
e-mail: leonard.nunney@ucr.edu

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.5230395.

THE ROYAL SOCIETY
PUBLISHING

# Determining cancer risk: the evolutionary multistage model or total stem cell divisions?

Leonard Nunney and Kevin Thai

Department of Evolution, Ecology and Organismal Biology, University of California, Riverside, CA 92521, USA

LN, 0000-0002-4315-3694

A recent hypothesis proposed that the total number of stem cell divisions in a tissue (TSCD model) determine its intrinsic cancer risk; however, a different model—the multistage model—has long been used to understand how cancer originates. Identifying the correct model has important implications for interpreting the frequency of cancers. Using worldwide cancer incidence data, we applied three tests to the TSCD model and an evolutionary multistage model of carcinogenesis (EMMC), a model in which cancer suppression is recognized as an evolving trait, with natural selection acting to suppress cancers causing a significant mean loss of Darwinian fitness. Each test supported the EMMC but contradicted the TSCD model. This outcome undermines results based on the TSCD model quantifying the relative importance of 'bad luck' (the random accumulation of somatic mutations) versus environmental and genetic factors in determining cancer incidence. Our testing supported the EMMC prediction that cancers of large rapidly dividing tissues predominate late in life. Another important prediction is that an indicator of recent oncogenic environmental change is an unusually high mean fitness loss due to cancer, rather than a high lifetime incidence. The evolutionary model also predicts that large and/or long-lived animals have evolved mechanisms of cancer suppression that may be of value in preventing or controlling human cancers.

## 1. Introduction

It has been argued that the total number of stem cell divisions occurring in a tissue is a critical indicator of cancer risk [1,2]. This total stem cell division (TSCD) model is supported by the strong log–log correlation between lifetime cancer risk (CR) and the lifetime number of stem cell divisions (LSCD) occurring in a tissue (the CR/LSCD correlation) evaluated across a number of cancers within the US ($r = 0.805$) [1] and worldwide (median $r = 0.80$) [2]. From this correlation, the TSCD model was used to suggest that about 65% of the differences in risk among cancers is due to the intrinsic effect of randomly occurring somatic mutations resulting from tissue-specific differences in LSCD. However, a different interpretation of the relationship using the same data indicated that the intrinsic effect of LSCD accounted for only 10–30% of cancer risk [3]. Regardless of this substantial difference in interpretation, a more fundamental question revolves around the validity of the TSCD model itself. Here we show that the high CR/LSCD regression/correlation is not robust and depends strongly on the types of cancers included. Additional tests using the same dataset used in the original research [1,2] fail to support the TSCD model; however, all tests support an alternative, the evolutionary model of multistage carcinogenesis (EMMC), a model that incorporates simple evolutionary principles into the traditional multistage model [4,5]. This model is based on the assumption that cancer suppression is an evolving trait and that the fitness loss due to the incidence of cancer in any given tissue in any given species is minimized by natural selection. As a result, the level (and hence genetics) of suppression potentially varies among tissues of the same species and potentially varies in the same tissue among different species. The EMMC resolves a serious problem with the traditional multistage model, which predicts that large long-lived animals (e.g. humans) should have a much higher incidence of cancer than small short-lived ones (e.g. mice) [6], yet they do not [7], a contradiction named Peto's paradox [4]. The support for the EMMC reinforces its value as an important tool in understanding patterns in the incidence of cancer across different tissues and different species.

The TSCD model has been widely criticized [8–12], and one of the recurring issues is that interpreting the CR/LSCD correlation in a causal fashion to infer intrinsic versus extrinsic causation is not consistent with the multistage model of carcinogenesis, a model that has formed the basis of our understanding of cancer risk since the 1950s [13,14]. At first sight, it might appear that the TSCD model is a simplification of the multistage model since, in both models, cancer initiation requires the accumulation of a cancer-specific set of driver mutations. However, there is a fundamental difference that makes the two models behave in profoundly different ways, as we will demonstrate. Under the multistage model, the set of necessary mutations must accumulate in a single cell, a pattern supported by detailed genomic sequencing [15,16], whereas we will show that the TSCD model implies that cancer initiation only requires that the necessary mutations accumulate across different cells anywhere within the tissue. In general, this would mean that no single-cell carries all of the necessary mutations, contrary to our well-established understanding of cancer initiation. In terms of modelling, this critical difference can be expressed in how the two models incorporate two of their major components: the number of stem cells in a tissue, $C$, and the total number of times each stem cell divides, $K$ (= $kt$, where $k$ is the probability of a cell dividing per unit time and $t$ defines age).

In the multistage model, $C$ and $K$ have different effects on cancer risk. A change in $C$ (cell number) has a linear (i.e. proportional) effect on cancer risk, since the chance of one cell in the tissue accumulating a specific set of mutations increases in proportion to the number of cells. On the other hand, the effect of a change in $K$, the total number of divisions per cell, is amplified by the number of driver mutations required ($M$, noting that $M > 1$), i.e. the chance of any single mutation occurring increases with $K$ so the chance of all $M$ independent mutations occurring within a specific cell lineage increases as $K^M$. Thus, given the most basic form of the multistage model, the risk of cancer ($p$) up to a given age (and given $p$ is small) can be defined by

$$p = C(Ku)^M, \tag{1.1}$$

where $u$ is the somatic mutation rate [4]. This basic model assumes that when any cell in the tissue has acquired the $M$ driver mutations it initiates cancer. The model can also approximate stepwise clonal expansion by assuming that the somatic mutation rate increases as driver mutations accumulate (since if the mutated cell expands by a factor $x$, then the probability of a mutation in any one of the clone also increases by $x$), so that the value of $u$ in equation (1.1) is the geometric mean of these sequential values [4].

By contrast, under the TSCD model, the total number of stem cell divisions accumulated by any time $t$ (=$TSCD(t)$ where LSCD is evaluated at $t = T$, the total lifespan) is $CK$, which in turn, it is suggested, predicts the incidence of cancer [1]. From equation (1.1), it can be seen that only in the unrealistic case of a single driver mutation initiating cancer ($M = 1$) does LSCD (= $CK$) incorporate the effect of $C$ and $K$ as defined in the multistage model. In general, the TSCD model can be expressed as

$$p = A(TSCD(t))^B = A(CK)^B = A(Ckt)^B \tag{1.2}$$

where $A$ and $B$ are constants (and recalling that $k$ is the rate of cell division per cell). The slope of the CR/LSCD regression (=$B$) for the US data (calculated from data used in ref. [1]) is 0.53.

In the TSCD model, $C$ and $K$ are assumed to act equivalently on the accumulation of driver mutations, and therefore, we can identify two extremes that have the same outcome under the TSCD model. It should not matter if a given set of oncogenic mutations arises within a single-cell lineage ($C = 1$) by increasing $K$ (which becomes increasingly probable under the multistage model) or across a tissue that only divides once ($K = 1$) by increasing $C$ (which remains very unlikely under the multistage model). This example illustrates our conclusion that the TSCD model does not incorporate the critical assumption of the multistage model that cancer is initiated when a set of $M$ driver mutations have accumulated in a single cell. Instead, it assumes that cancer is initiated once these driver mutations are present in any combination of 1 to $M$ cells anywhere within the tissue.

There is strong evidence for the role of cell number in driving an increased cancer risk across breeds of domestic dogs [17,18], and within humans the data are consistent with a linear relationship, i.e. $p \propto C$ [19]. This pattern supports the multistage model and may be consistent with the TSCD model, although the estimated value of $B = 0.53$ ($p \propto C^{0.53}$) defines a somewhat nonlinear effect (equation (1.2)). In any event, the predicted influence of $K$, i.e. the number of times a cell divides, is very different in the two models ($K^M$ versus $K^B$), given that typically $M > 2$ [15,16] while $B < 1$.

If TSCD is not the driving parameter of cancer risk, then why is the correlation between LSCD and lifetime cancer risk so high? Based on the best available estimates of LSCD for 31 categories of cancer using US data [1] and for 17 cancer types using global data [2], the CR/LSCD regression yields a correlation centred around $r = 0.80$. Given the multistage model, some correlation is expected since LSCD can be substituted into equation (1.1) and thereby eliminate either $C$ or $K$, but not both:

$$\begin{aligned}\ln(p) &= \ln(\mathrm{LCSD}) + [M\ln(u) + (M-1)\ln(K)]\\ &= M\ln(\mathrm{LCSD}) + [M\ln(u) - (M-1)\ln(C)].\end{aligned} \tag{1.3}$$

Thus, under this model and considering the conditions under which the terms within the square brackets are constant, the slope of the CR/LSCD regression should range from 1 (given a constant $M$ and $K$, with C varying among tissues) up to a slope of $M$ (given a constant $M$ and $C$, with only $K$ varying among tissues). As noted above, the slope for the US data is 0.53, a value that is substantially less than that expected based on the multistage model; however, it is well established that $M$ varies across cancer types [16,20]. Grouping cancers based on anatomical site, on the assumption that similar tissues would generally have similar $M$ and $K$ but different $C$, gave a within-group slope that was not significantly different from 1, in agreement with the multistage model [8], i.e. although the groups were assumed to differ in $M$ and $K$, within each of these groups the results were consistent with $p \propto C$ (see equation (1.1)).

This observed variation in $M$ is an important prediction of the evolutionary model of multistage carcinogenesis (EMMC). The model proposes that natural selection acts on the incidence of each cancer to minimize its effect on Darwinian fitness, so that if cancer causes a high level of mortality before reproduction is complete, then the loss of fitness will select for an increase the level of suppression acting on that cancer. More precisely, if cancer causes an average fitness loss greater than about $1/(2N_e)$, where $N_e$ is the effective population size, then selection will favour genetic variants with enhanced suppression of that cancer, typically resulting in an increase in $M$, the number of driver mutations required to initiate cancer [4]. When the average fitness loss across the population is smaller than $1/(2N_e)$, the selection is rarely effective [21].

## 2. Test 1: early-onset cancers and the CR/LSCD correlation

Most cancers are rare early in life but become much commoner in old age when their effect on Darwinian fitness is zero or nearly so. However, a few cancers are primarily pre-reproductive (e.g. paediatric cancers) and peak at an early age. These early-life cancers typically originate in tissues with stem cells that have very limited (or zero) divisions in adult life, plus these tissues usually have relatively small stem cell populations (e.g. retinoblastoma [20]), leading to a low LSCD. These cancers lack the massive late-life increase typical of most cancers, with the result that they have a relatively low lifetime incidence despite being more frequent than other cancers early in life. This raises the possibility that early life cancers have a disproportionate effect on the CR/LSCD correlation.

The TSCD model predicts that the magnitude of the correlation and the slope of the regression should not be affected by the presence or absence of early life cancers, since in this model, it does not matter when the cell divisions making up the lifetime total occur. On the other hand, the linkage of a low lifetime CR with a low LSCD typical of early-onset cancers could result in a CR/LSCD correlation driven in large part by early life cancers. This linkage is consistent with the EMMC, since natural selection is expected to act to limit (to the extent possible) cancers inexorably linked to early life events (such as growth) to a narrow time window, because these cancers have such a direct effect in reducing fitness. Thus, removing cancers that are primarily pre-reproductive from the analysis is expected to result in a much weaker correlation and a shallower slope under the EMMC but not under the TSCD model.

## 3. Test 2: early-onset cancers and the CR($t$)/TSCD($t$) correlation

A related but more rigorous prediction of the TSCD model can be examined by expanding the single-point CR/LSCD correlation (evaluated only using the oldest age class) into the curve of the CR($t$)/TSCD($t$) correlations over the whole lifetime, where the accumulated cancer risk up to any age $t$ (= CR($t$)) is plotted against the total stem cell divisions up to that age (= TSCD(t)) (again both being on a log scale). Letting $t = T$ gives lifetime values.

The TSCD model predicts that the CR($t$)/TSCD($t$) correlation should be independent of $t$, since the accumulation of stem cell divisions is hypothesized to drive cancer risk directly and therefore the age at which TSCD is evaluated should be irrelevant (see equation (1.2)), and hence the strong positive correlation estimated from lifetime data should be age-independent. On the other hand, the EMMC predicts that early-onset cancers will cause the correlation to decline and become negative as age is reduced. This pattern is expected because of a combination of two effects. First, typical late-life cancers are extremely rare early in life since at that time the probability of a single cell accumulating a complete set of driver mutations is very small. This is especially true in tissues with large rapidly dividing stem cell populations (e.g. the colon) which are expected to have the highest $M$. Thus, at early ages, their high TSCD is linked to a low CR. By contrast, at young ages, early-onset cancers are expected to have a low TSCD (due to a low division rate and small tissue size) but high CR (relative to cancers occurring primarily at older ages) associated with a low $M$ (e.g. retinoblastoma where $M = 2$ [20]). This logic is the basis of our test 2 and predicts a negative CR(t)/TSCD(t) correlation for low vales of $t$ that becomes positive as age (=$t$) increases. Second, there is a predicted evolutionary effect that is the basis of our test 3.

## 4. Test 3: late-onset cancers and the CR($t$)/TSCD($t$) correlation

If the potentially confounding effect of early-life cancers is removed, then it remains the case that the TSCD model predicts that the CR($t$)/TSCD($t$) correlation should be constant independent of age (= $t$). This is not the case under the EMMC model, since there is an additional evolutionary factor that is predicted to reduce the CR($t$)/TSCD($t$) correlation at young ages. Natural selection, acting to maintain the average fitness loss at no more than about $1/(2N_e)$, will tend to equalize the incidence of different cancers up to the end of the reproductive period. Thus, among the typical late-onset cancers, it is expected that when reproduction is more-or-less complete those cancers originating in large rapidly dividing tissues (i.e. those with high TSCD) will generally have a similar cumulative incidence to those originating in smaller, more slowly dividing tissues (i.e. those with low TSCD). However, cancer incidence in high TSCD tissues is predicted to increase relatively more rapidly during post-reproductive life, because in old age (when the equalizing effect of natural selection is absent), the more rapid accumulation of somatic mutations in high TSCD tissues will result in cancer rates that increase faster than those in low TSCD tissues [5]. By way of illustration, consider a tissue that has avoided cancer initiation at the onset of the post-reproductive age, but some fraction of its cells has $M - 1$ driver mutations. The rapidity with which cancer subsequently arises increases with the rate of cell division (which increases the somatic mutation rate per unit time) and the number of cells (which increases the probability of one cell becoming cancerous). This phenomenon predicts that, under EMMC, the CR(t)/TSCD(t) correlation will increase from around the end of reproduction and be strongest at the oldest ages, independent of the effect of juvenile cancers.

A more rigorous way of testing this expectation is to examine the change with age in the slope of the regression of log(cancer risk at $t$) versus log(stem cell divisions at $t$), evaluated across the adult cancers during a period $\Delta t$ ending at $t$. We used 5-year intervals ending at 20, 25, 30 years, etc. and since the data are not cumulative, all time periods are independent.

Given the TSCD model, the slope of this regression is expected to be the same regardless of the age at which it is evaluated, since differentiating equation (1.2) with respect to $t$ gives

$$\text{cancer risk at } t = \frac{\mathrm{d}p}{\mathrm{d}t} = AB(SCD)^{B}t^{B-1}, \tag{4.1}$$

where SCD (= $Ck$) is the number of stem cell divisions occurring in the tissue at $t$. Thus, across a group of tissues, the expectation is a linear relationship with a slope of $B$ between the log of cancer risk ($\log(\mathrm{d}p/\mathrm{d}t)$) in a given tissue during the interval $\Delta t$ and the log of the number of SCD in that tissue during the same time period, regardless of age. On the other hand, given the EMMC, the slope of this regression is expected to increase with age, most notably once the reproductive period comes to an end. This increase in slope is expected because, as outlined above, the cancer risk of high TSCD tissues accelerates relative to low TSCD tissues in older ages as reproduction, and hence fitness constraints, decline.

## 5. Material and methods

Following [2], the CI5-X (Cancer Incidence in Five Continents Volume X) 'detailed database', as provided by IARC at http://ci5.iarc.fr/CI5-X/Pages/download.aspx, was used for geographical age-specific cancer incidence, and using these data (for 2003–2007), we repeated their analysis of LSCD versus lifetime cancer risk of 17 cancers across 5 continents (Oceania, North America, Latin America/Caribbean, Asia and Europe), adopting their assumption that the oldest data (greater than 85 years) were well approximated by 90 years. As in [2], we also analysed the data from Africa, which ended at a maximum age of 80 years. Our analysis differed in that (i) the sexes were separated, which reduced the number of cancers from 17 overall to 15 per sex (since two were specific to each sex), and (ii) since there were two groupings of thyroid cancer and two of leukaemia originating from exactly the same tissue, to avoid potential bias due to subdividing the products of the same stem cells, these pairs were combined, giving 13 cancers per sex. The cancers (with the corresponding World Health Organization's ICD-10 code in parentheses) were head and neck (3 and 75), oesophageal squamous (24), colorectal carcinoma (42 and 49), hepatocellular carcinoma (59), pancreatic (70), lung adenocarcinoma (79), osteosarcoma (90), melanoma (100), breast (113, female only), ovarian germ cell (142, female only), prostate (151, male only), testicular (152, male only), medulloblastoma (189), thyroid (follicular 201 and papillary 202 combined with medullary 203) and leukaemia (chronic lymphocytic 229 combined with acute myeloid 233).

The total stem cell divisions at age $t$ (TSCD($t$)) were calculated from estimates of the number of stem cells ($C$) and of cell division cycles after growth ($K$) provided in [1], except that the prostate and oesophageal estimates were from [2], which also provided the estimate we used for the number of female breast stem cells; however, for the division rate of the breast stem cells we used cell turnover times from [22]: 22 days for age groups 15–25, 70 days for 30–35, 147 days for 40–90.

The CI5-X data includes 423 cancer registries spanning 68 different countries. The registries were pooled into the continents defined above. For countries with incidence data covering both the entire population and smaller regions, only the larger-scale cancer registry was used (e.g. Canada 2003–2007). For countries where incidence data did not include one for the entire population, the incidence rate was calculated by pooling the available regional registries (e.g. Argentina, Tierra del Fuego 2003–2007; Argentina, Mendoza 2003–2007; etc.) by summing the cases and the total person-years within each age group for each cancer type. The Hawaii data were removed from the USA dataset that was included in North America to prevent double representation since Hawaii was grouped in Oceania. Following [2], the data for Algeria (Setif), Malawi (Blantyre), South Africa (PROMEC), Iceland and Sweden were excluded as incidence data for the tissues they considered were unavailable. Furthermore, we removed all regions with incomplete person-year data (due to some regions listing the number of recorded cases but not listing the associated person-years in one or more of the older age classes). This resulted in 24 regions removed from Asia, 2 removed from Europe and 9 removed from Latin America. For a complete list of regions used for each continent, see electronic supplementary material, table S1.

Tomasetti & Vogelstein [1] estimated TSCD in tissue with $C$ stem cells and $K$ division cycles after growth as $C(2+K)-2$. This original formula contains a minor error. The correct formula is $\mathrm{TSCD} = C(1+K)-1$. For example, for a tissue to grow to 8 cells, there must be 7 (i.e. $C-1$) cell divisions, not 14 as indicated by the original formula. However, except when there are very few post-growth divisions, this change has a negligible effect.

## 6. Results

The TSCD model and the evolutionary model of multistage carcinogenesis were subjected to three tests using the same worldwide cancer risk data, sub-divided into continents, that was used in [2], except the data were separated by sex and there was a small correction in how the early-life growth phase was incorporated (see Material and methods). Of necessity, the same cancers were used due to the need for the estimates of stem cells and stem cell division rates that were compiled in [1,2]. Two pairs of cancers used in the original analyses that originate from the same tissue (two thyroid cancers and two leukaemias) were combined, resulting in the examination of 13 cancers per sex. Combining the two pairs of cancers slightly increased the overall mean correlation between the log of lifetime cancer risk and the log of the lifetime total number of stem cell divisions (the CR/LSCD correlation) of the pooled data (i.e. worldwide), but excluding the African data (since it does not include the two oldest age classes; see [2]), from an average across sexes of 0.802 to 0.811 (table 1).

## 7. Test 1: early-onset cancers and the CR/LSCD correlation

The dataset included two early-onset cancers seen in both sexes (medulloblastoma and osteosarcoma) and two sex-specific ones (ovarian and testicular germ cell cancer). When these cancers were excluded from the dataset, the overall mean CR/LSCD correlation (averaged across sexes) dropped to 0.444 and the slope declined, with the two sexes showing an almost identical pattern in the pooled data (figure 1). These changes were consistent across all 5 continent/sex combinations used in the pooled data, as well as in the data from Africa (table 1). The probability of all 12 independent tests showing the EMMC-predicted decreased correlation (due to a shallower slope) by chance is $p = 0.0002$ (sign test). Under the TSCD model, excluding juvenile cancers should not

**Table 1.** Worldwide CR/LSCD correlations with and without early-onset cancers, evaluated at 85+ years. The LSCD data are from [1,2], and, following them, 85+ is approximated as 90 years.

| | adult and early-onset | | | | adult only | |
|---|---|---|---|---|---|---|
| | 15 cancers[a] | | 13 cancers | | 10 cancers | |
| continent | female | male | female | male | female | male |
| N. America | 0.7970 | 0.7866 | 0.8037 | 0.7903 | 0.4169 | 0.4035 |
| Latin America/Caribbean | 0.7561 | 0.7017 | 0.7679 | 0.7108 | 0.3383 | 0.2306 |
| Europe | 0.8337 | 0.7942 | 0.8376 | 0.8007 | 0.5310 | 0.4548 |
| Asia | 0.6275 | 0.6367 | 0.6727 | 0.6606 | 0.0540 | 0.0482 |
| Oceania | 0.8201 | 0.8008 | 0.8300 | 0.8081 | 0.5091 | 0.4884 |
| Africa[b] | 0.7610 | 0.6649 | 0.7609 | 0.6611 | 0.3753 | 0.1174 |
| *pooled data*[c] | *0.8093* | *0.7940* | *0.8172* | *0.8052* | *0.4505* | *0.4369* |
| *significance of correlation*[c] | *0.0003* | *0.0004* | *0.0006* | *0.0009* | 0.1909 | 0.2068 |
| regression slope[c] | 0.430 | 0.520 | 0.423 | 0.493 | 0.219 | 0.223 |
| intercept[c] | −4.85 | −5.53 | −4.68 | −5.12 | −2.34 | −2.10 |

[a] includes two pairs of cancers from the same tissue that were combined in all other analyses.
[b] evaluated up to age 80.
[c] pooled data (all continents except Africa).

materially alter the correlation or slope; instead, in the pooled data, it resulted in a 70% relative drop in the variance explained, from 66% to 20%, turning highly a significant correlation ($p < 0.001$) into a non-significant one ($p > 0.1$) (table 1). However, this decrease is consistent with the EMMC.

## 8. Test 2: early-onset cancers and the CR($t$)/TSCD($t$) correlation

The CR($t$)/TSCD($t$) defines the relationship between the log of cancer risk up to age $t$ and the log of the number of accumulated stem cell divisions up to that age. Under the TSCD model, this relationship should be independent of age, hence the strong positive correlation seen when the lifetime data (i.e. $t = T$) are used (recalling $\mathrm{CR}(T)/\mathrm{TSCD}(T) \equiv \mathrm{CR}/\mathrm{LSCD}$) should be maintained regardless of age. By contrast, the EMMC predicts that the inclusion of early-onset cancers will cause the correlation to decline and become negative as age is reduced.

The negative correlation predicted by the EMMC was found in both sexes when the age was reduced to 20 years in the data from all five continents making up the pooled result (Asia, Europe, North America, Latin America/Caribbean and Oceania) and in the data from Africa (which only goes to age 80 years) (electronic supplementary material, table S2). In the pooled data, the correlation dropped from 0.818 and 0.805 at age 90 to −0.314 and −0.286 at age 20, in females and males, respectively (figure 2, solid curve).

## 9. Test 3: late-onset cancers and the CR($t$)/TSCD($t$) correlation

In the absence of the effect of juvenile cancers, the EMMC predicts that the CR($t$)/TSCD($t$) correlation will increase from around the end of reproduction and be strongest at the oldest ages. The alternative prediction, based on the TSCD model, is that the correlation will remain constant with age.

Analysis of the same worldwide dataset, with the three early-acting cancers in each sex removed, showed a clear reduction in the CR($t$)/TSCD($t$) correlation with age across all 6 × 2 continent × sex combinations (electronic supplementary material, table S2), which is a highly significant concordance ($p = 0.0002$, sign test). This pattern is shown for the pooled data (excluding Africa) in figure 2 (dashed line), where the correlation drops from 0.451 and 0.437 at age 90 to −0.008 and 0.149 at age 20, for females and males, respectively. This drop, predicted by the EMMC, violates the expectation of a constant correlation predicted by the TSCD model.

To get a clearer picture of this effect, it is possible to examine the change in the 'instantaneous' cancer risk (rather than the accumulated cancer risk) with age. This analysis has the advantage that each measure of cancer risk is independent (which is not true when using the accumulated risk). The TSCD expectation is of a linear relationship between the log of a given cancer's risk during a 5-year period and the log of the tissue's SCD during that period plotted across the 13 cancers, and that this relationship should be the same regardless of age. On the other hand, given the EMMC, the slope of this regression is expected to increase with age, most notably once the reproductive period comes to an end.

The pooled data strongly confirmed the EMMC expectation of a progressive increase in the regression slope with post-reproductive age. Plotting this slope at each age against age using the pooled data (figure 3) shows in a highly significant trend between age 40 and 90 years ($r = 0.974$ females, and 0.979 males, both $p < 0.00001$). This pattern was consistent across all 10 of the continent/sex combinations represented in the pooled data, with a very high post-reproductive period correlation in all cases ($r > 0.95$) (electronic supplementary material, table S3).

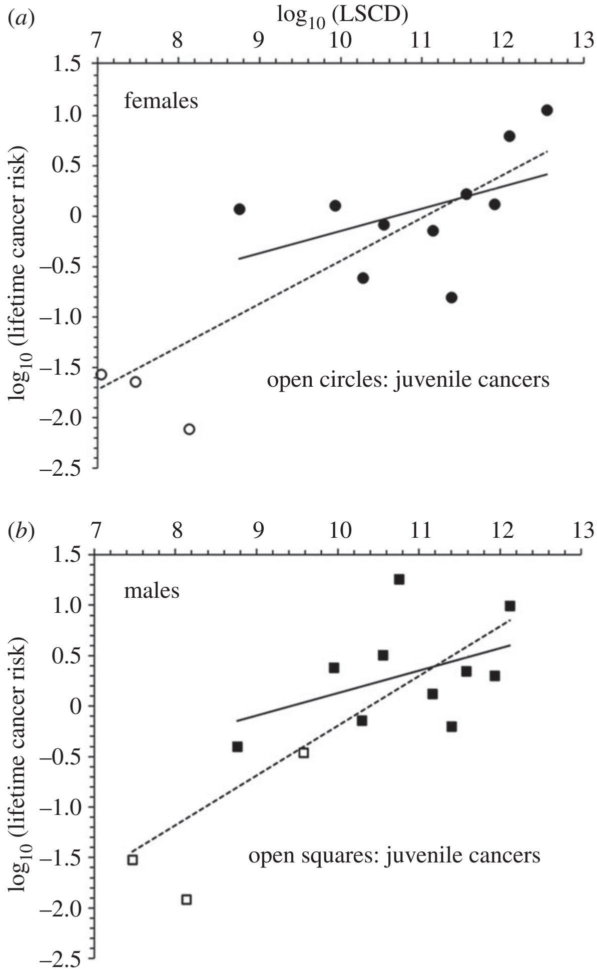

**Figure 1.** The CR/LSCD regression (= log(lifetime cancer risk) versus log(lifetime stem cell division)) based on the pooled worldwide lifetime incidence of 11 non-reproductive and two reproductive cancer types in females and in males. The regression of all data points (dashed line) is compared with the regression resulting after the three juvenile/early-adult cancer types were removed (solid line, filled symbols). If cancer risk was driven primarily by the lifetime number of stem cell divisions (TSCD model), the two regressions would have the same slope. The correlations for the pooled data and for individual continents are shown in table 1.

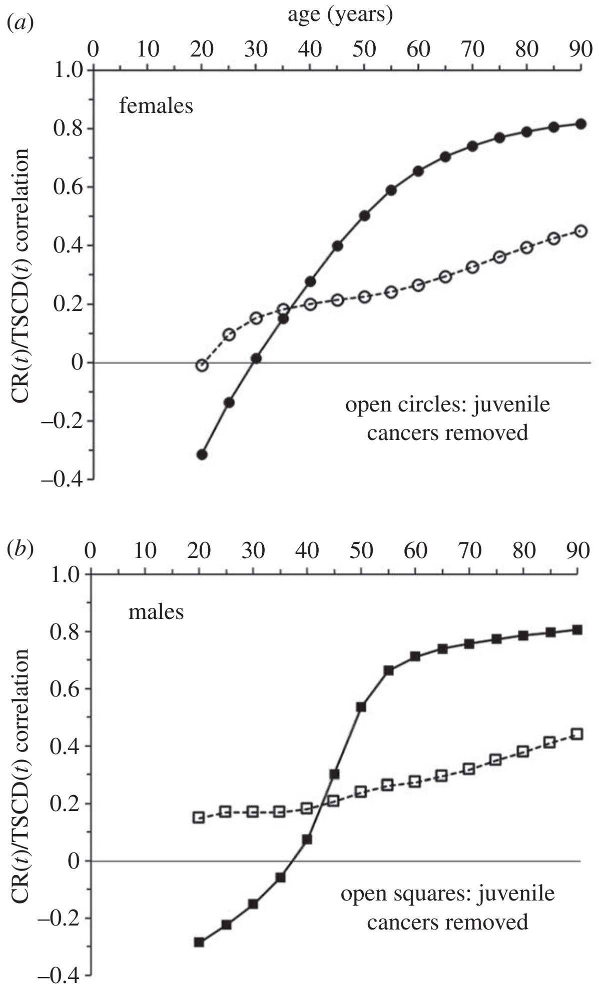

**Figure 2.** The change with age in the correlation between log(cancer risk to age $t$)) and log(stem cell division to age $t$), i.e. the CR($t$)/TSCD($t$) correlation. The curves shown used all 13 cancers (solid curve/filled symbols) or the 10 cancers remaining after the three juvenile/early-adult cancer types were removed (dashed line/open symbols). If cancer risk was driven primarily by the total number of stem cell divisions at a given age (TSCD model), the correlation would remain constant.

The only exception was the continent of Africa (measured from age 40 to the maximum age recorded of 80 years). Neither sex showed the robust pattern seen in all other continents. To understand this difference, we tested the hypothesis that the cross-sectional data used for the analysis encompassed some environmental changes in Africa that had relatively recently increased the incidence of some cancers, an effect that would be shown by a markedly lower cancer risk late in life relative to the other continents (the pooled data) not evident at the youngest ages. This pattern would be consistent with a change that did not affect older individuals, but was affecting younger ones, and it was seen for one non-reproductive cancer (lung adenocarcinoma) and one reproductive cancer (breast and prostate) in each sex, with incidence levels greater than 10 × lower in Africa than the pooled data at age 80, but with similar or higher incidence at ages 40 and 20. Removing them (leaving eight adult cancers) resulted in the pattern seen in the other continents, with correlation coefficients increasing from 0.26 to 0.90 (female) and 0.56 to 0.81 (male) (electronic supplementary material, table S3). Removing the same cancers from the pooled data maintained the same pattern, increasing the correlation coefficient from 0.974 and 0.979 to 0.997 and 0.998 in females and males, respectively.

## 10. Discussion

We used worldwide cancer risk data to test the hypothesis that cancer risk is driven by the total number of stem cell divisions occurring in the at-risk tissue (the TSCD model) [1,2]. In each case, the results were inconsistent with the TSCD model. By stark contrast, the evolutionary model of multistage carcinogenesis (EMMC) [4,5] was strongly supported.

The TSCD model has been used to attempt to define how much of our cancer burden is due to 'bad luck' (random but unavoidable somatic mutations), as distinguished from the effects of heredity and environmental factors (1–3). In the light of present results, conclusions drawn from the application of the TSCD model should be treated with caution. In any event, the term 'bad luck' is fraught with problems of definition since cancer is not a deterministic disease; in

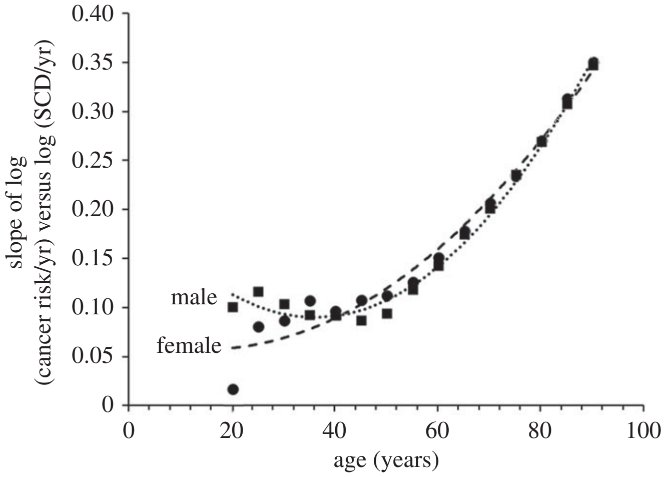

**Figure 3.** The increase in the slope of the regression of log(cancer risk at age *t*) on log(stem cell divisions at age *t*) for the 10 adult cancers evaluated over 5-year periods (i.e. the data are not cumulative, unlike the data in figure 2). If cancer risk was driven primarily by the total number of stem cell divisions occurring within a given period (TSCD model), the slopes would be independent of age and define a horizontal line (see equation (4.1)).

most cases, even for individuals experiencing an environmental risk factor (such as smoking), only a minority of those individuals succumb to relevant cancer and hence had, in some sense, bad luck [12,23,24]. But even accepting the intended premise, there is a major problem in partitioning out the three categories of risk (intrinsic, hereditary and environmental). Support of the EMMC reinforces the view that natural selection has acted to suppress early-onset cancer given the environment that humans experienced in the past. However, the effect of this selection on different types of cancer is expected to vary in a complicated fashion (see fig. 2 in [5]). Some cancers are expected to be rare due to very effective suppression, with most early-life cases being in individuals with inherited mutations and very few due to 'bad luck' in individuals with no germline mutation; however, other cancers are expected to be less suppressed with a frequency close to the threshold when natural selection is no longer effective, i.e. when their average fitness loss is about $1/(2N_e)$. In this latter case, most early-life cancers would arguably be due to 'bad luck' because the typical individual with cancer would have accumulated the initiating driver mutations somatically despite having the best available genotype [5].

Recognizing the existence of the $1/(2N_e)$ fitness threshold is important because it provides an understanding of why cancer, even early-onset cancer, is not completely eliminated. Once cancer is sufficiently rare (i.e. below a level causing the threshold fitness loss), then natural selection becomes ineffective at driving it to an even lower frequency [4]. On the other hand, a recent change in the environment can result in one or more cancers causing a larger than expected fitness loss because they are not in evolutionary equilibrium. Despite the expected variability in the evolutionary equilibrium level of suppression, cancer incidence should be at or below the threshold level. This knowledge can help in the search for conditions found in the modern environment that have increased the risk of some cancers. In some cases, we can speculate with some confidence on what such a factor might be. For example, the incidence of melanoma has steadily increased over the last 50 years in most fair-skinned populations in young as well as older age groups [25], reducing mean fitness. The movement of individuals from the more northerly parts of Europe to areas with much more intense UV radiation combined with behavioural changes are implicated. Similarly, there are good reasons to believe that the incidence of breast cancer has increased dramatically in recent times due to changes in nutrition and reproductive patterns, with estimates of the long-term increase in risk running as high as 100× [26].

Natural selection acts to minimize fitness loss, hence the EMMC makes fewer predictions regarding late-life cancer risk, since any effect on fitness is minimal. However, our test 3 strongly supported the EMMC prediction that cancers occurring in tissues with a large stem cell population and high division rate would show the highest increase in old age [5]. This result suggests an explanation for the intriguing observation of a marked increase in the proportion of epithelial cancers with age [27].

The importance of natural selection in modifying cancer risk is being increasingly recognized in the study of non-model animals, with the goal of revealing potentially useful mechanisms of cancer suppression that they have evolved [28–30]. As noted earlier, in the absence of evolution, the traditional multistage model incorrectly predicts that large and/or long-lived animals will have a much higher incidence of cancer (Peto's paradox). The solution to this paradox lies with the adaptive evolution built into the EMMC that maintains cancer risk at a low level regardless of changes in life history [4]. This model provides the best-supported explanation of Peto's paradox [31], and has prompted a focus on large mammals such as elephants [7,32] and whales [33,34], and on mammals with unusual life-history strategies, such as the long-lived naked mole rat [30]. It is very probable that such studies will ultimately lead to novel approaches to the cure or control of cancer in humans.

Data accessibility. Data are publicly available at: http://ci5.iarc.fr/CI5-X/Pages/download.aspx.

Authors' contributions. L.N. formulated the hypotheses, designed the research and analysed the data. L.N. and K.T. performed the research and wrote the paper.

Competing interests. The authors declare no competing interests.

Funding. We received no funding for this study.

## References

1. Tomasetti C, Vogelstein B. 2015 Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science* **347**, 78–81. (doi:10.1126/science.1260825)
2. Tomasetti C, Li L, Vogelstein B. 2017 Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science* **355**, 1330–1334. (doi:10.1126/science.aaf9011)
3. Wu S, Powers S, Zhu W, Hannun YA. 2016 Substantial contribution of extrinsic risk factors to cancer development. *Nature* **529**, 43–47. (doi:10.1038/nature16166)
4. Nunney, L. 1999 Lineage selection and the evolution of multistage carcinogenesis. *Proc. R. Soc. Lond. B* **266**, 493–498. (doi:10.1098/rspb.1999.0664)
5. Nunney L. 2003 The population genetics of multistage carcinogenesis. *Proc. R. Soc. Lond. B* **270**, 1183–1191. (doi:10.1098/rspb.2003.2351)
6. Peto R. 1977 Epidemiology, multistage models, and short-term mutagenicity tests. In *The origins of human cancer, vol. 4* (eds HH Hiatt, JD Watson, JA Winsten), pp. 1403–1428. New York, NY: Cold Spring Harbor Laboratory Press.
7. Abegglen LM *et al*. 2015 Potential mechanisms for cancer resistance in elephants and comparative cellular response to DNA damage in humans. *JAMA* **314**, 1850–1860. (doi:10.1001/jama.2015.13134)
8. Noble R, Kaltz O, Hochberg ME. 2015 Peto's paradox and human cancers. *Phil. Trans. R. Soc. B* **370**, 20150104. (doi:10.1098/rstb.2015.0104).
9. Nunney L, Muir B. 2015 Peto's paradox and the hallmarks of cancer: constructing an evolutionary framework for understanding the incidence of cancer. *Phil. Trans. R. Soc. B* **370**, 20150161. (doi:10.1098/rstb.2015.0161)
10. Rozhok AI, Wahl GM, DeGregori J. 2015 A critical examination of the 'bad luck' explanation of cancer risk. *Cancer Prev. Res.* **8**, 762–764. (doi:10.1158/1940-6207.CAPR-15-0229)
11. Noble R, Kaltz O, Nunney L, Hochberg ME. 2016 Overestimating the role of environment in cancers. *Cancer Prev. Res.* **9**, 773–776. (doi:10.1158/1940-6207.CAPR-16-0126)
12. Nowak MA, Waclaw B. 2017 Genes, environment, and 'bad luck'. *Science* **355**, 1266–1267. (doi:10.1126/science.aam9746)
13. Nordling CO. 1953 A new theory on the cancer-inducing mechanism. *Br. J. Cancer* **7**, 68–72. (doi:10.1038/bjc.1953.8)
14. Armitage P, Doll R. 1954 The age distribution of cancer and a multi-stage theory of carcinogenesis. *Br. J. Cancer* **8**, 1–12. (doi:10.1038/bjc.1954.1)
15. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz Jr LA, Kinzler KW. 2013 Cancer genome landscapes. *Science* **339**, 1546–1558. (doi:10.1126/science.1235122)
16. Martincorena I *et al*. 2017 Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1–13. (doi:10.1016/j.cell.2017.09.042)
17. Fleming J, Creevy K, Promislow D. 2011 Mortality in North American dogs from 1984 to 2004: an investigation into age-, size-, and breed-related causes of death. *J. Vet. Intern. Med.* **25**,187–198. (doi:10.1111/j.1939-1676.2011.0695.x)
18. Nunney L. 2013 The real war on cancer: the evolutionary dynamics of cancer suppression. *Evol. Applic.* **6**, 11–19. (doi:10.1111/eva.12018)
19. Nunney L. 2018 Size matters: height, cell number, and a person's risk of cancer. *Proc. R. Soc. B* **285**, 20181743. (doi:10.1098/rspb.2018.1743)
20. Knudson AG. 1971 Mutation and cancer: statistical study of retinoblastoma. *Proc. Natl Acad. Sci. USA* **68**, 820–823. (doi:10.1073/pnas.68.4.820)
21. Wright S. 1931 Evolution in Mendelian populations. *Genetics* **16**, 97–159.
22. Meyer JS. 1977 Cell proliferation in normal human breast ducts, fibroadenomas, and other ductal hyperplasias measured by nuclear labeling with tritiated thymidine: effects of menstrual phase, age, and oral contraceptive hormones. *Human Pathol.* **8**, 67–81. (doi:10.1016/S0046-8177(77)80066-X)
23. Weinberg CR, Zaykin D. 2015 Is bad luck the main cause of cancer? *JNCI* **107**, djv125. (doi:10.1093/jnci/djv125)
24. Nunney L. 2016 Commentary: the multistage model of carcinogenesis, Peto's paradox and evolution. *Int J. Epidemiol.* **43**, 649–653. (doi:10.1093/ije/dyv201)
25. Erdmann F *et al*. 2013 International trends in the incidence of malignant melanoma 1953–2008—are recent generations at higher or lower risk? *Int. J. Cancer.* **132**, 385–400. (doi:10.1002/ijc.27616)
26. Eaton SB *et al*. 1994 Women's reproductive cancers in evolutionary context. *Q. Rev. Biol.* **69**, 353–367. (doi:10.1086/418650)
27. DePinho RA. 2000 The age of cancer. *Nature* **408**, 248–254. (doi:10.1038/35041694)
28. Nunney L, Maley C, Breen M, Hochberg M, Schiffman J. 2015 Peto's paradox and the promise of comparative oncology. *Phil. Trans. R. Soc. B* **370**, 20140177. (doi:10.1098/rstb.2014.0177)
29. Tollis M, Schiffman JD, Boddy AM. 2017 Evolution of cancer suppression as revealed by mammalian comparative genomics. *Curr. Opinion Genet. Develop.* **42**, 40–47. (doi:10.1016/j.gde.2016.12.004)
30. Seluanov A, Gladyshev VN, Vijg J, Gorbunova V. 2018 Mechanisms of cancer resistance in long- lived mammals. *Nat. Rev. Cancer* **18**, 433–441. (doi:10.1038/s41568-018-0004-9)
31. Nunney L. 2020 Resolving Peto's paradox: modeling the potential effects of size-related metabolic changes, and of the evolution of immune policing and cancer suppression. *Evol Applic.* **13**, 1581–1592. (doi:10.1111/eva.12993)
32. Sulak M *et al*. 2016 TP53 copy number expansion is associated with the evolution of increased body size and an enhanced DNA damage response in elephants. *eLife* **5**, e11994. (doi:10.7554/eLife.11994)
33. Keane M *et al*. 2016 Insights into the evolution of longevity from the bowhead whale genome. *Cell Reports* **10**, 112–122. (doi:10.1016/j.celrep.2014.12.008)
34. Tollis M *et al*. 2019 Return to the sea, get huge, beat cancer: an analysis of cetacean genomes including an assembly for the humpback whale (*Megaptera novaeangliae*). *Mol. Biol. Evol.* **36**, 1746–1763. (doi:10.1093/molbev/msz099)